

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE NOV 2013		2. REPORT TYPE		3. DATES COVERED 00-00-2013 to 00-00-2013	
4. TITLE AND SUBTITLE Structured Knowledge Space				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Massachusetts Institute of Technology, Lincoln Laboratory, 244 Wood Street, Lexington, MA, 02420-9108				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as Report (SAR)	18. NUMBER OF PAGES 2	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

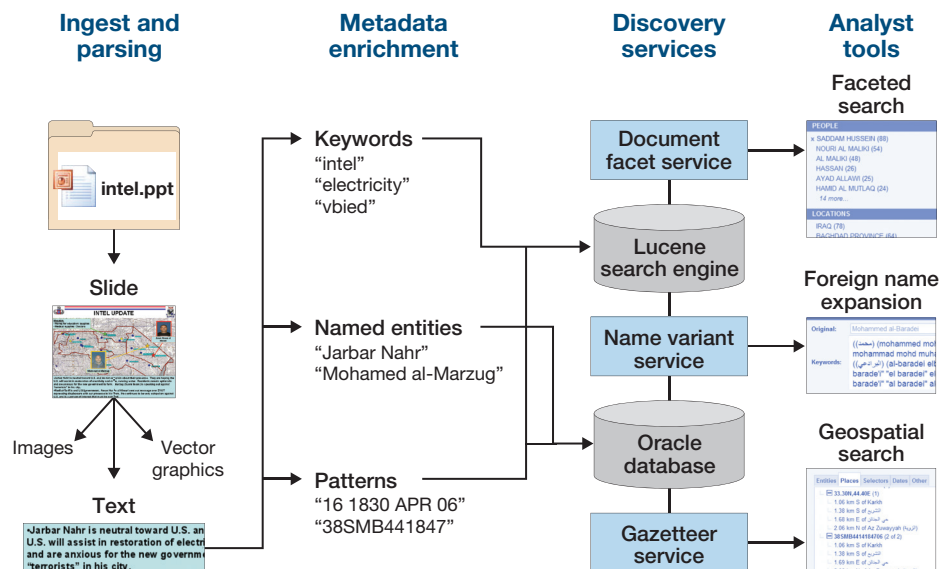
Structured Knowledge Space

A multifaceted software system enables increased exploitation of a vast store of intelligence and military reporting.

Structured Knowledge Space (SKS) is an end-to-end software system developed to solve a problem that has frustrated national security decision makers: “How do we take advantage of the enormous amounts of information communicated daily through a wide variety of reporting venues?” Various factors make it difficult for decision makers to search and correlate the wealth of information contained in these reports:

- Documents are often stored in Microsoft PowerPoint, Adobe PDF, or other formats not well suited to search or to computer-based analysis.
- Reports are often disseminated via email or other ad hoc channels, further hindering search and discovery of critical battlefield or intelligence information.
- The number and variety of organizations involved leads to significant volume and velocity of data lacking a coordinated indexing system.
- Although documents vary greatly, from brief daily updates to lengthy analyses, they commonly use domain-specific jargon and abbreviations; “boilerplate” text, such as headers and disclaimers, provide no new information but clog the search process.

SKS combines open-source technologies (e.g., Java and Lucene), custom-built software, and domain knowledge about important entities in intelligence reporting to create a robust system that



Structured Knowledge Space (SKS) creates structured metadata (essentially data about other data) to improve the discovery and use of unstructured reports, i.e., reports such as Word documents or email that are not organized in a predefined model such as a database or table.

facilitates searching over a document collection that had previously been largely unsearchable. SKS builds searchable archives of text-based intelligence reports, extracts information from free-form documents, and makes the information discoverable through a keyword and faceted-search interface. SKS's tools include ones that search for approximate name matches or geographic locations referenced in text. SKS's modern tiered architecture scales to significant data storage and retrieval demands.

SKS exploits modern natural language processing and information retrieval techniques to improve the ability to search, analyze, and effectively utilize intelligence reports and the valuable information that they contain. Its functionality is similar to

niche capabilities in other industries, e.g., Google News for aggregating news sources and Radian6 for social media analysis. However, SKS was designed to meet the specific needs of the military and intelligence communities.

Capabilities of SKS

SKS started as an R&D effort and has since been productized and fully integrated into several customer information processing and dissemination chains. SKS's features increase users' capability to exploit the knowledge captured in the multitude of intelligence and operational documents generated and filed each day:

- Users can query for approximate name matches or geographic locations referenced in documents.

- Special features deal with transliterated Arabic names, which present challenges because of the inconsistent spellings that arise when Arabic characters are represented with English letters. This capability was driven by specific user needs that made more general-purpose commercial software less useful.
- SKS includes features for data browsing and trend analysis.

SKS's functionality relies on its ability to efficiently and accurately recognize and extract entities from documents. An entity is the textual representation of a person's name, possibly including military rank; an organization's name; a place name (city, region, country, etc.); or specialized entities such as date-time groups (a common way of representing dates and times in the U.S. military) and geospatial coordinates. SKS employs rules and dictionaries to enable discovery and extraction of such entities. Several of the rule-based extractors are quite complex, so in order to make them more computationally efficient, they are implemented as tries, i.e., tree data structures for efficient retrieval of words and phrases.

Extracted entities are indexed to enable efficient search and discovery. SKS creates structured metadata (essentially data about other data, e.g., source of the data, date the data were collected, size of a data file) to improve indexing. This indexing is also based on various similarity scores, thus allowing users to search by exact match or to search for documents similar to ones already discovered. SKS is capable of searching for documents containing a geospatial coordinate or time reference within a specified geospatial or temporal region of interest. The system also provides a reverse gazetteer, which describes where extracted geospatial coordinates are located relative to named locations. SKS can also search by ingest date and by word or phrase trends, thus improving analysts' ability to connect related information.

Data discovery and extraction by SKS go beyond the individual-document level by offering capabilities for summarizing data holdings at the result set and corpus-wide levels. While some systems can show counts of entities or phrases across multiple documents, SKS provides analysts with summaries of key topics across

The screenshot shows the SKS search interface with several annotations in red boxes:

- Search box:** Standard Google-like search operators
- Advanced search:** Search by geo, source, ingest date, reported date, document type
- Facet categories:** Entities listed by document count; selecting an entity adds it to current search
- Search results:** Results ranked by relevance; search terms are highlighted to show context
- Find similar documents:** The [Similar] link is a new search for all documents that contain similar text
- Preview and download a document**
- Plot geocoordinates for one or all search results on a visualization tool (e.g., GoogleEarth)**

The Structured Knowledge Space search page provides diverse, useful information.

the whole corpus. Thus, SKS enables users to view data at the level of detail appropriate for their current task.

SKS includes techniques for clustering documents into groups with similar content. This capability allows users to rapidly scan topics available in a document collection to help them find the subset of most interest. SKS's flexible mechanisms for ingesting documents include an upload web page and the ability to monitor email accounts and directories. Because of this flexibility, SKS can be used as a general-purpose document repository and discovery tool, e.g., on a company intranet.

Benefits of SKS

SKS's features increase military and intelligence analysts' ability to make use of the large collection of documents generated each day. As an illustration of the scope of data SKS can handle, a feed of information from the Open Source Center generated approximately 3000 new documents per day for an SKS development system.

SKS offers a service that did not previously exist. SKS can perform document-clustering that reveals connections that may be extremely useful to analysts by

- Finding all documents referring to an organization (even when the organization has several aliases and/or name variations)
- Finding all documents referring to a particular person (even when the per-

son has several aliases and/or name transliterations)

- Finding all documents with a geospatial reference within a certain distance of a location
- Finding all documents with a time reference within a specified date range

SKS is providing a much needed capability in the national security domain. The current users are primarily the military and intelligence communities; however, other communities, such as law enforcement or border protection, may find use for information gleaned from the reporting. The near-term road map for SKS includes increasing the sophistication of its text-mining algorithms and providing early demonstrations of unstructured data processing on the Department of Defense's emerging cloud platforms. ■

Technical Point of Contact

Delsey Sherrill
Intelligence & Decision
Technologies Group
dsherrill@ll.mit.edu
781-981-4699

For further information, contact

Communications and
Community Outreach Office
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420-9108
781-981-4204